Project 3: Analyses of daily COVID-19 cases across nations

Jiayi Shen (js5354), Siquan Wang (sw3442), Jack Yan (xy2395)

5/1/2020

Introduction

The pandemic of COVID-19 is the biggest challenge that the world is facing right now. Our lifes are all deeply affected by this public health crisis. By building a model on the growth of COVID-19 cases, we can have a better understanding of the current status and then plan future responses. Thus analyzing existing data and predicting future trajectories has become the most important task faced by public health expertises and policy makers.

The objective of this project includings:
- Develop an optmization algorithm to fit a logisitc curve to each region, using the global COVID-19 data;
- Evaluate how appropiate it is to use such model;
- Apply clustering methods to the logistic grwoth model parameters, and observe the patterns.

Data

The datasets used in this project are adapted from https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports and it recorded the following variables:

Id: Record ID

Province/State: The lcoal state/province of the record; 54% records do not have this info;

Country/Region: The country/regionoof the record;

Lat: Lattudiute of the record;

Long: Longitude of the record;

Date: Date of the record; from Jan 21 to March 23;

ConfirmedCases: The number of confirme case on that day;

Fatalities: The number of death on that day;

For the purpose of fitting logistic growth curve, the main variables of interest are `Country/Region` and `ConfirmedCases`. We groupped the dataset by `Country/Region`, and pulled out `days since first case` and `cumulative confirmed cases on each day`.

Method

Method: Logisitic curves

Logisitic curves could be one way to model the trajectory of cumulative cases; It is a parametric function with the form

$$f(t) = \frac{a}{1 + exp\{-b(t-c)\}},$$

where t is the days since the first infection; a is the upper bound, i.e. the maxium number of cases a region can reach, b is growth rate, and c is the mid-point, where the curve changes from convex to concave; Each curve is uniquely defined by (a, b, c). By design a logistic curve increases exponentially at begining and slows down at the end.

Consider the most commonly used loss function: Mean Squared Error (MSE), which takes the form

$$l = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \frac{a}{1+\exp\{-b(T_i-c)\}})^2$$

Then our goal is to minimize the above loss function with respect to (a, b, c). To do this, we can apply Newton-Raphson algorithm. Let $\boldsymbol{\theta} = (a, b, c)$. Newton's method suggests to update $\boldsymbol{\theta}$ iteratively, such that the ith step is given by

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} - [\nabla^2 l(\boldsymbol{\theta}_{i-1})]^{-1}\nabla l(\boldsymbol{\theta}_{i-1})$$

where $\nabla l(\boldsymbol{\theta}_{i-1})$ is the gradient, and $[\nabla^2 l(\boldsymbol{\theta}_{i-1})]^{-1}$ is the Hessian matrix.

In this particular case, we have incorporated some of the technique introduced in 'Newton's method with a large p' lecture session. To be more specific, we have replaced the Hessian matrix with an identity matrix to simplify the computation and increase the efficiency. Then to make sure the algorithm will converge correctly, step-halfing is also incorporated to control the step size for the monotone trend. Then the ith step of our Newton algorithm is given by

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} + \lambda \boldsymbol{H}_{i-1,p\times p}\nabla l(\boldsymbol{\theta}_{i-1})$$

where $\boldsymbol{H}_{i-1,p\times p} = I_{p\times p}$, $\lambda \in (0, 1)$.

The gradient vector $\nabla l(\boldsymbol{\theta}_{i-1})$ of our loss function is:

$$\begin{pmatrix} \partial l/\partial a \\ \partial l/\partial b \\ \partial l/\partial c \end{pmatrix} = \begin{pmatrix} \frac{2}{n}\sum_{i=1}^{n}(Y_i - \frac{a}{1+\exp\{-b(T_i-c)\}}) \cdot \frac{-1}{1+\exp\{-b(T_i-c)\}} \\ \frac{2}{n}\sum_{i=1}^{n}(Y_i - \frac{a}{1+\exp\{-b(T_i-c)\}}) \cdot \frac{-a(T_i-c)\exp\{-b(T_i-c)\})}{(1+\exp\{-b(T_i-c)\})^2} \\ \frac{2}{n}\sum_{i=1}^{n}(Y_i - \frac{a}{1+\exp\{-b(T_i-c)\}}) \cdot \frac{ab\exp\{-b(T_i-c)\})}{(1+\exp\{-b(T_i-c)\})^2} \end{pmatrix} \quad (1)$$

We set the convergence criteria to be that the difference in MSE of two consecutive iterations is smaller than 10^{-5}, and the maximum iteration number to be 1000.

After setting up the above Newton's method, we applied our algorithm to the whole dataset and summarized out result in 'res_df'. How should we report this result? (eg, visualization, pick some typical countries, summary table)

Method: Clustering

Results

Task 1.1

Develop an optmization algorithm to fit a logisitc curve to each region, and find a way to visualize your fitted curves effectively; What you learn from your fitted models? e.g. how many regions have passed the midpoint? how many regions are approaching to the end of the virus speading; Which regions have faster growth rates and which regions have more "flat" growth?

Task 1.2

You can find daily reports after March 23 from the following github site

https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports

From those data, do you think if the logistic curve is a reasonable model for fitting the curmulative casesa and predicting future new cases?

I think logistic curve might not be an appropriate model for fitting the curmulative cases and predicting future new cases. First our Newton's algorithm suggests that although we spent much efforts in optimizing this optimization algorithm (including standardization, replacing Hessian by Identity mattrix and step-halfing), the convergence speed might still have some unknown internal issue and the results highly depending on the initial starting value if we do not do standardization, which means that our model is not that robust and the estimated curve shape parameters might be questionable in some unusual case. Second, some African countries' results are little bit strange base on our model, which might be due to too liitle data available. Finally, I may suggest using some other exponential-distribution based curves to fit the data and incorporate more shape paramters in the model fitting process. Or we could some hybrid approaches to fit different types of curves based on preclassification of countries.

Task 2: Clustering your fitted curves

clustering is an effective data exploring tools; It helps develop hypothesis and identify potential risk factors; Apply K-mean and Guassian mixture model (with EM algorithm) to cluster the fitted parameters $(\hat{a}, \hat{b}, \hat{c})$; Whicl algoirhtm does a better job in clustering those curves? Are the resulting clusters related to geogrpahic regions, or the starting timeing of the local virus speading, or the resources of the regions? You may use external informations to help understand the clusters, i.e. find plausible explanations why some regions have similar (a, b, c)?